The Retrievability of Information

Leif Azzopardi
School of Computing Science, University of Glasgow
Glasgow, United Kingdom
Leif.Azzopardi @glasgow.ac.uk

ABSTRACT

Retrievability is an important and interesting indicator that can be used in a number of ways to analyse Information Retrieval systems and document collections. Rather than focusing totally on relevance, retrievability examines what is retrieved, how often it is retrieved, and whether a user is likely to retrieve it or not. This is important because a document needs to be retrieved, before it can be judged for relevance. In this tutorial, we shall explain the concept of retrievability along with a number of retrievability measures, how it can be estimated and how it can be used for analysis. Since retrieval precedes relevance, we shall also provide an overview of how retrievability relates to effectiveness - describing some of the insights that researchers have discovered thus far. We shall also show how retrievability relates to efficiency, and how the theory of retrievability can be used to improve both effectiveness and efficiency. Then we shall provide an overview of the different applications of retrievability such as Search Engine Bias, Corpus Profiling, etc., before wrapping up with challenges and opportunities. The final session of the day will look at example problems and ways to analyse and apply retrievability to other problems and domains. Participants are invited to bring their own problems to create a more focused practical session. This tutorial is ideal for: (i) researchers curious about retrievability and wanting to see how it can impact their research, (ii) researchers who would like to expand their set of analysis techniques, and/or (iii) researchers who would like to use retrievability to perform their own analysis.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software:Performance Evaluation

General Terms

Theory, Experimentation

Keywords

Retrievability, Effectiveness, Evaluation, Simulation

1. INTRODUCTION

The tutorial will be broken into five main parts: (i) Definition, Theory and Measures of Retrievability (ii) The Estimation of Document Retrievability, (iii) The Relationship between Retrievability and Effectiveness, (iv) Applications of Retrievability and (v) Applying Retrievability to your own research. Finally, we will conclude with a summary of the challenges and directions of future research.

1.1 Definitions and Measures (1-1.5hrs)

In this part of the tutorial, we will introduce the different "abilities" in IR, which affect how findable a document is, either from a user's perspective or a system's perspective. And importantly how they related and are dependent upon each other. For example,a document needs to be indexed before it can be retrieved, and to be indexed a document needs to be crawled, etc. Once this context is set, we shall explicitly define retrievability explaining how it can be derived from first principles and how it can be derived through an analogy with Transportation Planning. This session is then concluded with an introduction to the different measurements that can be obtained. The following topics will be covered:

1. The -abilities of Information Retrieval
   - Findability [30]
   - Navigability [41, 21, 23]
   - Accessibility [28, 11]
   - Searchability [35]
   - Crawlability [29]
   - Discoverability [22]
   - Usability [32]
   - Retrievability [13]
2. What is retrievability? How easily can a document be found? What is the probability of finding a document? [13]
3. How Information Retrieval relates Transportation Planning [11]
   - Transportation planning, Land Use and Accessibility Measures [27, 26]

- Information Spaces vs. Physical Spaces
- An analogy between IR and Transportation Planning

4. Measure of Retrievability
   - Cumulative based measures
   - Gravity based measures
5. Retrievability Bias
   - The Lorenze Curve [25]
   - Gini Co-efficient [25]
   - Other inequality measure (Hoover, Theil, Palma, etc)

1.2 Estimating Retrievability (1hr)

The second part of the tutorial will focus on how to estimate the retrievability of a particular document (i.e. a page-centric approach) and how the retrievability of all the documents can be estimated (i.e. a collection-centric approach). The pragmatic problems of obtaining such estimates shall be discussed along with how to generate/simulate the queries required to formulate a reasonable estimate (thus we shall provide an overview of various ways to simulate queries). The type of estimate depends on how the measure will be used - so we shall describe how depending on the type of analysis to be performed which estimation techniques will be more appropriate. Also, we shall describe retrievability can be efficiently estimated depending on how it will be used. A summary of the topics we shall cover here are:

1. Page-Centric estimates versus Collection-Centric Estimates
2. The Universe of all Possible Queries
3. Absolute estimates versus Relative estimates
4. Generating Queries to estimate retrievability [5, 6, 1]
   - Single Term
   - Bigram / Biterm
   - n-grams
   - Title based
   - Query Logs
5. Efficient estimations and approximation of retrievability and bias
6. Relationship between Cumulative and Gravity based measures
7. Relationship between inequality measures
8. Analysis of documents and collections using Retrievability and how document collections can be analysed using retrievability? [12, 15, 17]

1.3 Relationships with Retrievability (1hr)

Part three will focus in on the relationship between retrievability and various retrieval effectiveness measures. Here we will describe and discuss the various efforts that have tried to understand the relationship between retrievability and effectiveness [19, 18, 40, 4, 39]. Firstly, from a theoretical point of view, we will discuss the different possible relationships and how retrievability can impact upon both effectiveness and efficiency. Then we will consider the relationship with effectiveness between different retrieval models (i.e. how well do the retrievability scores of various systems correlate to system effectiveness or system rankings) and within a retrieval model (i.e. how well does the retrievability of a particular retrieval model, given its different parameter settings, correlate to system effectiveness). Specifically, we shall show how retrievability can be used to rank systems and tune retrieval models.

1. The conceptual / hypothesised relationship between retrievability and performance [4]
2. Retrievability and Efficiency
3. The empirical relationship between retrievability [4, 39] and:
   - Mean Average Precision,
   - Precision and MRR,
   - Recall,
   - NDCG, etc.
4. The empirical relationship between retrievability and new user oriented gain based measures [38]
5. Retrievability and Retrieval Models when ranking systems [20]

1.4 Applications of Retrievability (1hr)

Part four will describe the research conducted by a growing number of research groups who have applied retrievability, or the theory of, to gain improvements in effectiveness and/or efficiency, or to gain other insights. Some of the research directions covered will include:

1. Search Engine Bias: how systems influence user populations [9, 37, 31].
2. Improving Recall: the highs and lows of affect retrievable patents [14, 19].
3. The Reverted Index: how retrievability turns retrieval on its head to produce improvements in both effectiveness and efficiency [33].
4. Psuedo Relevance Bias: how Pseudo Relevance is biased, and addressing that bias leads to performance improvements [18].
5. Findability: games that make you find while measuring how easily documents can be found [10, 34]. As part of this session, participants will be invited to play test out the games developed to measure how easily people can find pages given a search engine.

To wrap up, we will then outline the future directions and research challenges associated with retrievability. There are numerous research opportunities in how to use and apply retrievability research, as well as more basic research in terms of the estimation, relationships, theory and applications of retrievability.

1.5 Practical Session (1-2hrs)

The final part of the tutorial will be dedicated to questions and answers about retrievability, and going through in groups, how to apply retrievability in the domains that are of interest to the participants. In fact, participants will be invited to bring along and share their research problems and during the course of this session focus on designing a retrievability analysis along with the necessary experiments to undertake such an analysis.

2. INTENDED LEARNING OUTCOMES

By the end of the tutorial, students should be able to:

- Define and describe retrievability and retrievability bias
- Explain the relationship between retrievability, accessibility, findability, navigability, and other -abilities.
- Estimate the retrievability of documents within a collection
- Design a retrievability experiment to detect/monitor retrieval bias
- Describe the relationships between retrievability, effectiveness and efficiency
- Evaluate an application using retrievability measures

Students will be provided with handouts and the reference lists (in print and electronic from). Also, available online is code to compute retrievability scores (in Python and C++) [7].

3. PREREQUISITES, AUDIENCE, LENGTH

A basic understanding of Information Retrieval will be assumed. That is, we expect the attendees to know what a IR system is, the inputs and outputs of the system, the standard ways to evaluate a retrieval system, along with some knowledge of the different types of retrieval models (i.e. vector space, best match, etc) [36].
The intended audience would be students undertaking a PhD, or about to, and researchers that are interested in finding out about the recent developments that have been made regarding retrievability and how it can be used.
The length of the tutorial will be a day.

4. BIOGRAPHY

Leif Azzopardi is a Senior Lecturer within the School of Computing Science at the University of Glasgow, unofficial leader of the legendary Glasgow Information Retrieval Group, and pioneer of the theory of retrievability. His research focuses on building formal models for Information Retrieval - usually drawing upon different disciplines for inspiration, such as Quantum Mechanics, Operations Research, Microeconomics, Transportation Planning and Gamification. Central to his research is the theoretical development of statistical language models for Information Retrieval, where his research interests include:

- Models for the retrieval of documents, sentences, experts and other information objects [16, 24];
- Probabilistic models of user interaction and the simulation of users for evaluation [1, 5, 6];
- Microeconomic models of information interaction, specifically how cost and effort affect interaction and performance with search systems [2];
- Methods which assess the impact of search technology on society in application areas such as, search engine bias and the accessibility of e-Government information [13], and;
- Search for fun (i.e. the SINS of users) [3].

He has given numerous invited talks on Retrievability throughout the world and lectures at the Information Foraging Summer School (2011, 2012 and 2013) and Symposium of Future Directions in Information Access (2007-2013). He recently released a free online book called, *How to Tango with Django* which is a noob's guide to web development in Python's Django (available free at: **www.tangowithdjango.com** [8]).

He received his Ph.D. in Computing Science from the University of Paisley in 2006, and he received a First Class Honours Degree in Information Science from the University of Newcastle, Australia, 2001. In 2010, he received a Post-Graduate Certificate in Academic Practice and has been lecturing at the University of Glasgow since then.

5. REFERENCES

[1] L. Azzopardi. Query side evaluation: an empirical analysis of effectiveness and effort. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 556–563. ACM, 2009.

[2] L. Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 15–24, 2011.

[3] L. Azzopardi. Searching for unlawful carnal knowledge. In *Proceedings of the SIGIR Workshop: Search for Fun*, volume 11, pages 17–18, 2011.

[4] L. Azzopardi and R. Bache. On the relationship between effectiveness and accessibility. In *Proc. of the 33rd international ACM SIGIR*, pages 889–890, 2010.

[5] L. Azzopardi and M. de Rijke. Automatic construction of known-item finding test beds. In *Proceedings of SIGIR '06*, pages 603–604, 2006.

[6] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 455–462. ACM, 2007.

[7] L. Azzopardi, R. English, C. Wilkie, and D. Maxwell. Page retrievability calculator. In *To appear in the Proceedings of the European Conference in Information Retrieval*, ECIR '2014, 2014.

[8] L. Azzopardi and D. Maxwell. Tango with django: a begginners guide to web development in python / django, 2013.

[9] L. Azzopardi and C. Owens. Search engine predilection towards news media providers. In *Proc. of the 32nd ACM SIGIR*, pages 774–775, 2009.

[10] L. Azzopardi, J. Purvis, and R. Glassey. Pagefetch: a retrieval game for children (and adults). In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1010–1010, 2012.

[11] L. Azzopardi and V. Vinay. Accessibility in information retrieval. In *Proc. of the 30th ECIR*, pages 482–489, 2008.

[12] L. Azzopardi and V. Vinay. Document accessibility: Evaluating the access afforded to a document by the retrieval system. *Workshop on Novel Methodologies for Evaluation in Information Retrieval*, pages 52–60, 2008.

[13] L. Azzopardi and V. Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *Proc. of the 17th ACM CIKM*, pages 561–570, 2008.

[14] R. Bache. Measuring and improving access to the corpus. In *Current Challenges in Patent Information Retrieval*, volume 29 of *The Information Retrieval Series*, pages 147–165. 2011.

[15] R. Bache and L. Azzopardi. Transactions on large-scale data- and knowledge-centered systems ii. chapter Improving access to large patent corpora, pages 103–121. Springer-Verlag, 2010.

[16] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 43–50, 2006.

[17] S. Bashir and A. Rauber. Analyzing document retrievability in patent retrieval settings. In *Proceedings of the 20th International Conference on Database and Expert Systems Applications*, DEXA '09, pages 753–760. Springer-Verlag, 2009.

[18] S. Bashir and A. Rauber. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proc. of the 18th ACM CIKM*, pages 1863–1866, 2009.

[19] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *Proc. of the 32nd ECIR*, pages 457–470, 2010.

[20] S. Bashir and A. Rauber. On the relationship bw query characteristics and ir functions retrieval bias. *J. Am. Soc. Inf. Sci. Technol.*, 62(8):1515–1532, 2011.

[21] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 490–497. ACM, 2001.

[22] A. Dasgupta, A. Ghosh, R. Kumar, C. Olston, S. Pandey, and A. Tomkins. The discoverability of the web. In *Proc. of the 16th ACM WWW*, pages 421–430, 2007.

[23] X. Fang, P. Hu, M. Chau, H.-F. Hu, Z. Yang, and O. Sheng. A data-driven approach to measure web site navigability. *J. Manage. Inf. Syst.*, 29(2):173–212, Oct. 2012.

[24] R. T. Fernández, D. E. Losada, and L. A. Azzopardi. Extending the language modeling framework for sentence retrieval to include local context. *Information Retrieval*, 14(4):355–389, 2011.

[25] J. L. Gastwirth. The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, 54:306–316, 1972.

[26] S. L. Handy and N. D. A. Measuring accessibility: An exploration of issues and alternatives. *Environemnet and Planning A*, 29(7):1175–1194, 1997.

[27] W. Hansen. How accessibility shape land use. *Journal of the American Institute of Planners*, 25(2):73–76, 1959.

[28] S. Lawrence and L. Giles. Accessibility of information on the web. *Nature*, 400:101–107, 1999.

[29] A. Marchetto, R. Tiella, P. Tonella, N. Alshahwan, and M. Harman. Crawlability metrics for automated web testing. *International Journal on Software Tools for Technology Transfer*, pages 131–149, 2011.

[30] P. Morville. *Ambient Findability: What We Find Changes Who We Become*. O'Reilly Media, Inc., 2005.

[31] A. Mowshowitz and A. Kawaguchi. Assessing bias in search engines. *Information Processing and Management*, pages 141 – 156, 2002.

[32] J. W. Palmer. Web site usability, design, and performance metrics. *Info. Sys. Research*, 13(2):151–167, June 2002.

[33] J. Pickens, M. Cooper, and G. Golovchinsky. Reverted indexing for feedback and expansion. In *Proc. of the 19th ACM CIKM*, pages 1049–1058, 2010.

[34] J. Purvis and L. Azzopardi. A preliminary study using pagefetch to examine the searching ability of children and adults. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 262–265, 2012.

[35] T. Upstill, N. Craswell, and D. Hawking. Buying bestsellers online: A case study in search & searchability. In *7th Australasian Document Computing Symposium*, Sydney, Australia, 2002.

[36] C. J. van Rijsbergen. *Information Retrieval*. 1979.

[37] L. Vaughan and M. Thelwall. Search engine coverage bias: evidence and possible causes. *Information Processing and Management*, pages 693 – 707, 2004.

[38] C. Wilkie and L. Azzopardi. An initial investigation on the relationship between usage and findability. In *Advances in Information Retrieval*, pages 808–811. Springer, 2013.

[39] C. Wilkie and L. Azzopardi. Relating retrievability, performance and length. In *Proc. of the 36th ACM SIGIR conference*, SIGIR '13, pages 937–940, 2013.

[40] C. Wilkie and L. Azzopardi. Best and fairest: an empirical analysis of retrieval system bias. In *To appear in the Proceedings of the European Conference in Information Retrieval*, ECIR '2014, 2014.

[41] Y. Zhang, H. Zhu, and S. Greenwood. Web site complexity metrics for measuring navigability. In *Proc. of the 4th QSIC.*, pages 172–179, 2004.